Chiang et al. *Journal of Inequalities and Applications* 2013, 2013:580
http://www.journalofinequalitiesandapplications.com/content/2013/1/580

Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

Perturbation analysis of the stochastic algebraic Riccati equation

Chun-Yueh Chiang[1], Hung-Yuan Fan[2], Matthew M Lin[3*] and Hsin-An Chen[3]

*Correspondence:
mhlin@ccu.edu.tw
[3]Department of Mathematics, National Chung Cheng University, Chia-Yi 621, Taiwan
Full list of author information is available at the end of the article

Abstract

In this paper we study a general class of stochastic algebraic Riccati equations (SARE) arising from the indefinite linear quadratic control and stochastic H_∞ problems. Using the Brouwer fixed point theorem, we provide sufficient conditions for the existence of a stabilizing solution of the perturbed SARE. We obtain a theoretical perturbation bound for measuring accurately the relative error in the exact solution of the SARE. Moreover, we slightly modify the condition theory developed by Rice and provide explicit expressions of the condition number with respect to the stabilizing solution of the SARE. A numerical example is applied to illustrate the sharpness of the perturbation bound and its correspondence with the condition number.

MSC: Primary 15A24; 65F35; secondary 47H10; 47H14

Keywords: Brouwer fixed-point theorem; perturbation bound; stochastic algebraic Riccati equations; condition number

1 Introduction

In this paper we consider a general class of continuous-time stochastic algebraic Riccati equations

$$A^\top X + XA + C^\top XC - \big(XB + C^\top XD + S\big)\big(R + D^\top XD\big)^{-1}\big(B^\top X + D^\top XC + S^\top\big) + H = 0, \tag{1a}$$

$$R + D^\top XD \succ 0, \tag{1b}$$

where $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $D \in \mathbb{R}^{n\times m}$, $S \in \mathbb{R}^{n\times m}$, respectively. Moreover, $H \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are symmetric matrices. Here we denote $M \succ 0$ (respectively, $M \succeq 0$) if M is symmetric positive definite (respectively, positive semidefinite). The unknown $X \in \mathbb{R}^{n\times n}$ is a symmetric solution to SARE (1a)-(1b). Let $\mathcal{S}^n$ be the set of all symmetric $n \times n$ real matrices. For any $X, Y \in \mathcal{S}^n$, we write $X \succeq Y$ if $X - Y \succeq 0$.

In essence, SARE (1a)-(1b) is a rational Riccati-type matrix equation associated with the operator $\mathcal{R} : \operatorname{dom}\mathcal{R} \to \mathcal{S}^n$

$$\mathcal{R}(X) = \mathcal{P}(X) - \mathcal{S}(X)\mathcal{Q}(X)^{-1}\mathcal{S}(X)^\top,$$

where the affine linear operators $\mathcal{P} : \mathcal{S}^n \to \mathcal{S}^n$, $\mathcal{Q} : \mathcal{S}^n \to \mathcal{S}^m$, $\mathcal{S} : \mathcal{S}^n \to \mathbb{R}^{n\times m}$, and $\operatorname{dom}\mathcal{R}$ are defined by

$$\mathcal{P}(X) = A^\top X + XA + C^\top XC + H,$$

$$\mathcal{Q}(X) = R + D^\top XD,$$

$$\mathcal{S}(X) = XB + C^\top XD + S,$$

$$\operatorname{dom}\mathcal{R} = \left\{X \in \mathcal{S}^n \mid \mathcal{Q}(X) \succ 0\right\}.$$

We say that X is the maximal solution (or the greatest solution) of SARE (1a)-(1b) if it satisfies (1a)-(1b) and $X \succeq P$ for any $P \in \mathcal{S}^n$ satisfying $\mathcal{R}(P) \geq 0$ and (1b), *i.e.*, X is the maximal solution of $\mathcal{R}(X) \geq 0$ with the constraint (1b). Furthermore, it is easily seen that SARE (1a)-(1b) also contains the continuous-time algebraic Riccati equation (CARE)

$$A^\top X + XA - XBR^{-1}B^\top X + H = 0 \tag{2}$$

with $R \succ 0$, $C = 0$, $D = 0$ and $S = 0$, and the discrete-time algebraic Riccati equation (DARE)

$$X - C^\top XC + \left(C^\top XD + S\right)\left(R + D^\top XD\right)^{-1}\left(D^\top XC + S^\top\right) - H = 0 \tag{3}$$

with $A = \frac{-I}{2}$ and $B = 0$, as special cases.

Matrix equations of the type (1a)-(1b) are encountered in the *indefinite* linear quadratic (LQ) control problem [1], and the disturbance attenuation problem, which is in deterministic case the H_∞ control theory, for linear stochastic systems with both state- and input-dependent white noise. For example, see [2–4]. For simplicity, we only consider one-dimensional Wiener process of white noise in this paper; it is straightforward but tedious to extend all perturbation results presented in this paper for multi-dimensional cases. In the aforementioned applications of linear stochastic systems, a symmetric solution X, called a *stabilizing solution*, to SARE (1a)-(1b) ought to be determined for the design of optimal controllers. This stabilizing solution plays a very important role in many applications of linear system control theory. The definition of a stabilizing solution to SARE (1a)-(1b) is given as follows. (See also [3, Definition 5.2].)

Definition 1.1 Let $X \in \mathcal{S}^n$ be a solution to SARE (1a)-(1b), $\Phi = A + BF$ and $\Psi = C + DF$, where $F = -\mathcal{Q}(X)^{-1}\mathcal{S}(X)^\top$. The matrix X is called a stabilizing solution for $\mathcal{R}$ if the spectrum of the associated operator $\mathcal{L}_c$ with respect to X defined by

$$\mathcal{L}_c(W) = \Phi^\top W + W\Phi + \Psi^\top W\Psi, \quad W \in \mathcal{S}^n, \tag{4}$$

is contained in the open left half plane, *i.e.*, $\sigma(\mathcal{L}_c) \subset \mathbb{C}_-$.

Note that if $C = D = 0$ in (1a)-(1b), then it is easily seen from Definition 1.1 that the matrix $X \in \mathcal{S}^n$ is a stabilizing solution to SARE (1a)-(1b) or, equivalently, CARE (2) if and only if $\sigma(\Phi) \subset \mathbb{C}_-$. Therefore, Definition 1.1 is a natural generalization of the definition of a stabilizing solution to CARE (2) in classical linear control theory. Moreover, a necessary and sufficient condition for the existence of the stabilizing solution to a more general SARE is derived in Theorem 7.2 of [3]. See also [1, Theorem 10]. In this case, it is also shown that if SARE (1a)-(1b) has a stabilizing solution $X \in \operatorname{dom}(\mathcal{R})$, then it is necessarily a maximal solution and thus unique [1, 3].

The standard CARE (2) and DARE (3) are widely studied and play very important roles in both classical LQ and H_∞ control problems for deterministic linear systems [5–7]. In the past four decades, an extensive amount of numerical methods were studied and developed for solving the CARE and DARE (see [8–10] and the references therein). There are two major methodologies among these numerical methods or algorithms. One is the so-called *Schur method* or *invariant subspace method*, which was first proposed by Laub [11]. According to this methodology, the unique and non-negative definite stabilizing solution of the CARE (or DARE) can be obtained by computing the stable invariant subspace (or deflating subspace) of the associated Hamiltonian matrix (or symplectic matrix pencil). Some variants of the invariant subspace method, which preserve the structure of the Hamiltonian matrix (or symplectic matrix pencil) by special orthogonal transformations in the whole computational process, are considered by Mehrmann and his coauthors [12–18]. The other methodology comes from the *iterative method*, for example, it is referred to as Newton's method [6], matrix sign function method [19], disk function method [20], and structured doubling algorithms [21, 22] and references therein. So far there has been no sources in applying the invariant subspace methods for solving SARE (1a)-(1b), since the structures of associated Hamiltonian matrix or symplectic matrix pencil are not available. Only the iterative methods, *e.g.*, Newton's method [3] and the interior-point algorithm presented in [1], can be applied to computing the numerical solutions of SARE (1a)-(1b). Recently, normwise residual bounds were proposed for assessing the accuracy of a computed solution to SARE (1a)-(1b) [23].

Due to the effect of roundoff errors or the measurement errors of experimental data, small perturbations are often incorporated in the coefficient matrices of SARE (1a)-(1b), and hence we obtain the perturbed SARE

$$\begin{aligned}&\widetilde{A}^\top\widetilde{X}+\widetilde{X}\widetilde{A}+\widetilde{C}^\top\widetilde{X}\widetilde{C}-\big(\widetilde{X}\widetilde{B}+\widetilde{C}^\top\widetilde{X}\widetilde{D}+\widetilde{S}\big)\big(\widetilde{R}+\widetilde{D}^\top\widetilde{X}\widetilde{D}\big)^{-1}\big(\widetilde{B}^\top\widetilde{X}+\widetilde{D}^\top\widetilde{X}\widetilde{C}+\widetilde{S}^\top\big)\\&\quad+\widetilde{H}=0,\end{aligned}\tag{5a}$$

$$\widetilde{R}+\widetilde{D}^\top\widetilde{X}\widetilde{D}\succ 0,\tag{5b}$$

where $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{D}$, $\widetilde{H}$, $\widetilde{R}$ and $\widetilde{S}$ are perturbed coefficient matrices of compatible sizes. The main question is under what conditions perturbed SARE (5a)-(5b) still has a stabilizing solution $\widetilde{X}\in\mathcal{S}^n$. Moreover, how sensitive is the stabilizing solution $X\in\operatorname{dom}(\mathcal{R})$ of original SARE (1a)-(1b) with respect to small changes in the coefficient matrices? This is related to the conditioning of SARE (1a)-(1b). Therefore, we will try to answer these questions for SARE (1a)-(1b) in this paper. For CARE (2) and DARE (3), the normwise non-local and local perturbation bounds have been widely studied in the literature. See, *e.g.*, [24–26]. Also, computable residual bounds were derived for measuring the accuracy of a computed solution to CARE (2) and DARE (3), respectively [27, 28]. To our best knowledge, these issues have not been taken into account for constrained SARE (1a)-(1b) in the literature.

To facilitate our discussion, we use $\|\cdot\|_F$ to denote the Frobenius norm and $\|\cdot\|$ to denote the operator norm induced by the Frobenius norm. For $A=(A_1,\dots,A_n)=(a_{ij})\in\mathbb{R}^{m\times n}$ and $B\in\mathbb{R}^{p\times q}$, the Kronecker product of A and B is defined by $A\otimes B=(a_{ij}B)\in\mathbb{R}^{mp\times nq}$, and the operator $\operatorname{vec}(A)$ is denoted by $\operatorname{vec}(A)=(A_1^\top,\dots,A_n^\top)^\top$. It is known that

$$\operatorname{vec}(ABC)=\big(C^\top\otimes A\big)\operatorname{vec}(B),\qquad \operatorname{vec}\big(A^\top\big)=P_{n,m}\operatorname{vec}(A),$$

where $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{m\times \ell}$, $C \in \mathbb{R}^{\ell\times k}$, and $P_{n,m}$ is the Kronecker permutation matrix which maps $\mathrm{vec}(A)$ into $\mathrm{vec}(A^\top)$ for a rectangle matrix A, *i.e.*,

$$P_{n,m} = \sum_{i,j=1}^{n,m} E_{i,j,n\times m} \otimes E_{j,i,m\times n},$$

where the $n \times m$ matrix $E_{i,j,n\times m}$ has 1 as its (i,j) entry and 0's elsewhere.

This paper is organized as follows. In Section 2, a perturbation equation is derived from SAREs (1a)-(1b) and (5a)-(5b) without dropping any higher-order terms. By using Brouwer fixed point theorem, we obtain a perturbation bound for the stabilizing solution of SARE (5a)-(5b) in Section 3. In order to guarantee the existence of the stabilizing solution of perturbed SARE (5a)-(5b), some stability analysis of the operator $\mathcal{L}_c$ is established in Section 4. A theoretical formula of the normwise condition number of the stabilizing solution to SARE (1a)-(1b) is derived in Section 5. Finally, in Section 6, a numerical example is given to illustrate the sharpness and tightness of our perturbation bounds, and Section 7 concludes the paper.

2 Perturbation equation

Assume that $X \in \mathcal{S}^n$ is the unique stabilizing solution to SARE (1a)-(1b) and $\widetilde{X} \in \mathcal{S}^n$ is a symmetric solution of perturbed SARE (5a)-(5b), that is,

$$\mathcal{R}(X) := A^\top X + XA + C^\top XC - \Xi(X) + H = 0, \tag{6}$$

$$\widetilde{\mathcal{R}}(\widetilde{X}) := \widetilde{A}^\top \widetilde{X} + \widetilde{X}\widetilde{A} + \widetilde{C}^\top \widetilde{X}\widetilde{C} - \widetilde{\Xi}(\widetilde{X}) + \widetilde{H} = 0, \tag{7}$$

where the two operator $\Xi : \mathcal{S}^n \to \mathcal{S}^n$ and $\widetilde{\Xi} : \mathcal{S}^n \to \mathcal{S}^n$ are given by

$$\begin{aligned} \Xi(X) &= \mathcal{S}(X)\mathcal{Q}(X)^{-1}\mathcal{S}(X)^\top, \\ \widetilde{\Xi}(\widetilde{X}) &= \widetilde{\mathcal{S}}(\widetilde{X})\widetilde{\mathcal{Q}}(\widetilde{X})^{-1}\widetilde{\mathcal{S}}(\widetilde{X})^\top, \end{aligned} \tag{8}$$

and two affine linear operators $\widetilde{\mathcal{S}} : \mathcal{S}^n \to \mathcal{S}^n$, $\widetilde{\mathcal{Q}} : \mathcal{S}^n \to \mathcal{S}^m$ are defined by

$$\begin{aligned} \widetilde{\mathcal{S}}(\widetilde{X}) &= \widetilde{X}\widetilde{B} + \widetilde{C}^\top \widetilde{X}\widetilde{D} + \widetilde{S}, \\ \widetilde{\mathcal{Q}}(\widetilde{X}) &= \widetilde{R} + \widetilde{D}^\top \widetilde{X}\widetilde{D} \end{aligned}$$

for all $\widetilde{X} \in \mathcal{S}^n$. Let

$$\Delta X = \widetilde{X} - X.$$

The purpose of this section is to derive a perturbation equation of ΔX from SAREs (1a)-(1b) and (5a)-(5b). For the sake of perturbation analysis, we adopt the following notations:

$$\begin{aligned} &\Delta A = \widetilde{A} - A, \qquad \Delta B = \widetilde{B} - B, \qquad \Delta C = \widetilde{C} - C, \qquad \Delta D = \widetilde{D} - D, \\ &\Delta S = \widetilde{S} - S, \qquad \Delta R = \widetilde{R} - R, \qquad \Delta H = \widetilde{H} - H \end{aligned} \tag{9}$$

and

$$\begin{aligned} \delta Q &= \Delta R + D^\top X\Delta D + \Delta D^\top XD + \Delta D^\top X\Delta D, \\ \delta S &= \Delta S + C^\top X\Delta D + \Delta C^\top XD + \Delta C^\top X\Delta D + X\Delta B. \end{aligned} \tag{10}$$

Moreover, let

$$\begin{aligned} &F = -\mathcal{Q}(X)^{-1}\mathcal{S}(X)^{\top}, \qquad \widetilde{F} = -\widetilde{\mathcal{Q}}(X)^{-1}\widetilde{\mathcal{S}}(X)^{\top}, \\ &\Psi = C + DF, \qquad \widetilde{\Psi} = \widetilde{C} + \widetilde{D}\widetilde{F}, \\ &\Phi = A + BF, \qquad \widetilde{\Phi} = \widetilde{A} + \widetilde{B}\widetilde{F}, \end{aligned} \tag{11}$$

and by the definition of Ψ, we define

$$K := X\Psi. \tag{12}$$

Note that $\widetilde{\mathcal{S}}(X) = \mathcal{S}(X) + \delta S$ and $\widetilde{\mathcal{Q}}(X) = \mathcal{Q}(X) + \delta Q$. Substituting (11) into (8), we observe that

$$\begin{aligned} &\Xi(X) = -\mathcal{S}(X)F, \\ &\widetilde{\Xi}(\widetilde{X}) = \bigl(\widetilde{\mathcal{S}}(X) + \Delta X\widetilde{B} + \widetilde{C}^{\top}\Delta X\widetilde{D}\bigr)\bigl(\widetilde{\mathcal{Q}}(X) + \widetilde{D}^{\top}\Delta X\widetilde{D}\bigr)^{-1} \\ &\qquad \times \bigl(\widetilde{\mathcal{S}}(X) + \Delta X\widetilde{B} + \widetilde{C}^{\top}\Delta X\widetilde{D}\bigr)^{\top}. \end{aligned} \tag{13}$$

Thus far, we have not specified the relation between $\mathcal{R}(X)$ and $\widetilde{\mathcal{R}}(\widetilde{X})$. Such a tedious task can be turned into a breeze by repeatedly applying the matrix identities [29]

$$(I + U)^{-1} = I - U(I + U)^{-1}, \qquad V(I + UV)^{-1} = (I + VU)^{-1}V. \tag{14}$$

To begin with, assume that ΔR and ΔD are sufficiently small so that $\widetilde{\mathcal{Q}}(X)$ is invertible. We see that the product

$$\begin{aligned} &\bigl(\widetilde{\mathcal{S}}(X) + \Delta X\widetilde{B} + \widetilde{C}^{\top}\Delta X\widetilde{D}\bigr)\bigl(\widetilde{\mathcal{Q}}(X) + \widetilde{D}^{\top}\Delta X\widetilde{D}\bigr)^{-1} \\ &\quad = \bigl(\widetilde{\mathcal{S}}(X) + \Delta X\widetilde{B} + \widetilde{C}^{\top}\Delta X\widetilde{D}\bigr) \\ &\qquad \times \bigl[I - \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\bigl(I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\bigr)^{-1}\bigr]\widetilde{\mathcal{Q}}(X)^{-1} \\ &\quad = -\widetilde{F}^{\top} + \Delta X\widetilde{B}\bigl(I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\bigr)^{-1}\widetilde{\mathcal{Q}}(X)^{-1} \\ &\qquad + \widetilde{\Psi}^{\top}\Delta X\widetilde{D}\bigl(I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\bigr)^{-1}\widetilde{\mathcal{Q}}(X)^{-1}. \end{aligned}$$

It follows that

$$\begin{aligned} \widetilde{\Xi}(\widetilde{X}) = {} & -\widetilde{\mathcal{S}}(X)\widetilde{F} - \widetilde{F}^{\top}\widetilde{B}^{\top}\Delta X - \widetilde{F}^{\top}\widetilde{D}^{\top}\Delta X\widetilde{C} - \widetilde{\Psi}^{\top}\Delta X\widetilde{D}\widetilde{F} - \Delta X\widetilde{B}\widetilde{F} \\ & + \bigl(\widetilde{\Psi}^{\top}\Delta X\widetilde{D} + \Delta X\widetilde{B}\bigr)\bigl(I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\bigr)^{-1}\widetilde{\mathcal{Q}}(X)^{-1} \\ & \times \bigl(\widetilde{D}^{\top}\Delta X\widetilde{\Psi} + \widetilde{B}^{\top}\Delta X\bigr) \end{aligned}$$

since $\widetilde{F}^{\top}\widetilde{\mathcal{S}}(X)^{\top} = \widetilde{\mathcal{S}}(X)\widetilde{F}$. Next, from (11) we can see that

$$\begin{aligned} &\widetilde{\Phi}^{\top}\Delta X + \Delta X\widetilde{\Phi} = \widetilde{A}^{\top}\Delta X + \Delta X\widetilde{A} + \widetilde{F}^{\top}\widetilde{B}^{\top}\Delta X + \Delta X\widetilde{B}\widetilde{F}, \\ &\widetilde{\Psi}^{\top}\Delta X\widetilde{\Psi} = \widetilde{C}^{\top}\Delta X\widetilde{C} + \widetilde{F}^{\top}\widetilde{D}^{\top}\Delta X\widetilde{C} + \widetilde{\Psi}^{\top}\Delta X\widetilde{D}\widetilde{F}. \end{aligned} \tag{15}$$

Applying (15), we obtain the linear equation

$$\widetilde{\mathcal{R}}(\widetilde{X}) - \mathcal{R}(X) = \widetilde{\Phi}^{\top}\Delta X + \Delta X\widetilde{\Phi} + \widetilde{\Psi}^{\top}\Delta X\widetilde{\Psi} - E - h_2(\Delta X) = 0, \tag{16}$$

where

$$\begin{aligned}
E &:= -\big(\Delta A^{\top}X + X\Delta A + \widetilde{C}^{\top}X\widetilde{C} - C^{\top}XC + \widetilde{\mathcal{S}}(X)\widetilde{F} - \mathcal{S}(X)F + \Delta H\big),\\
h_2(\Delta X) &:= \widetilde{\Psi}^{\top}\Delta X\widetilde{D}\big(I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\big)^{-1}\widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{\Psi}\\
&\quad + \Delta X\widetilde{B}\big(I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\big)^{-1}\widetilde{\mathcal{Q}}(X)^{-1}\widetilde{B}^{\top}\Delta X\\
&\quad + \Delta X\widetilde{B}\big(I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\big)^{-1}\widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{\Psi}\\
&\quad + \widetilde{\Psi}^{\top}\Delta X\widetilde{D}\big(I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D}\big)^{-1}\widetilde{\mathcal{Q}}(X)^{-1}\widetilde{B}^{\top}\Delta X.
\end{aligned}$$

It follows from (16) that

$$\widetilde{\Phi}^{\top}\Delta X + \Delta X\widetilde{\Phi} + \widetilde{\Psi}^{\top}\Delta X\widetilde{\Psi} = E + h_2(\Delta X). \tag{17}$$

Equipped with this fact, we now are going to derive a perturbation equation in terms of ΔX by using ΔA, ΔB, ΔC, ΔD, ΔS, ΔR, δS, and δQ. It should be noted that

$$\begin{aligned}
\widetilde{\Psi} &= (\Delta C + C) - (\Delta D + D)\big(\mathcal{Q}(X) + \delta Q\big)^{-1}\big(\mathcal{S}(X) + \delta S\big)^{\top}\\
&= \Psi + \Delta\Psi,
\end{aligned}$$

with

$$\begin{aligned}
\Delta\Psi &:= \Delta C - \Delta D\mathcal{Q}(X)^{-1}\mathcal{S}(X)^{\top} - \Delta D\mathcal{Q}(X)^{-1}\delta S^{\top} - D\mathcal{Q}(X)^{-1}\delta S^{\top}\\
&\quad + (\Delta D + D)\mathcal{Q}(X)^{-1}\delta Q\mathcal{Q}(X)^{-1}\big(I + \mathcal{Q}(X)^{-1}\delta Q\big)^{-1}\big(\mathcal{S}(X)^{\top} + \delta S^{\top}\big)
\end{aligned} \tag{18}$$

and

$$\begin{aligned}
\widetilde{\Phi} &= (\Delta A + A) - (\Delta B + B)\big(\mathcal{Q}(X) + \delta Q\big)^{-1}\big(\mathcal{S}(X) + \delta S\big)^{\top}\\
&= \Phi + \Delta\Phi,
\end{aligned}$$

with

$$\begin{aligned}
\Delta\Phi &:= \Delta A - \Delta B\mathcal{Q}(X)^{-1}\mathcal{S}(X)^{\top} - \Delta B\mathcal{Q}(X)^{-1}\delta S^{\top} - B\mathcal{Q}(X)^{-1}\delta S^{\top}\\
&\quad + (\Delta B + B)\mathcal{Q}(X)^{-1}\delta Q\mathcal{Q}(X)^{-1}\big(I + \mathcal{Q}(X)^{-1}\delta Q\big)^{-1}\big(\mathcal{S}(X)^{\top} + \delta S^{\top}\big).
\end{aligned} \tag{19}$$

It then is natural to express the left-hand side of (17) by $\Delta\Phi$ and $\Delta\Psi$ such that

$$\widetilde{\Phi}^{\top}\Delta X + \Delta X\widetilde{\Phi} + \widetilde{\Psi}^{\top}\Delta X\widetilde{\Psi} = \Phi^{\top}\Delta X + \Delta X\Phi + \Psi^{\top}\Delta X\Psi - h_1(\Delta X),$$

with

$$h_1(\Delta X) := -\big(\Delta\Phi^{\top}\Delta X + \Delta X\Delta\Phi + \Psi^{\top}\Delta X\Delta\Psi + \Delta\Psi^{\top}\Delta X\Psi + \Delta\Psi^{\top}\Delta X\Delta\Psi\big).$$

Observe further that

$$\begin{aligned}
\widetilde{C}^{\top}X\widetilde{C} - C^{\top}XC &= C^{\top}X\Delta C + \Delta C^{\top}XC + \Delta C^{\top}X\Delta C,\\
\widetilde{S}(X)\widetilde{F} - \mathcal{S}(X)F &= -\big(\mathcal{S}(X)+\delta S\big)\big(\mathcal{Q}(X)+\delta Q\big)^{-1}\big(\mathcal{S}(X)+\delta S\big)^{\top}\\
&\quad + \mathcal{S}(X)\mathcal{Q}(X)^{-1}\mathcal{S}(X)^{\top}\\
&= F^{\top}\delta S^{\top} + \delta SF - \delta S\big(I+\mathcal{Q}(X)^{-1}\delta Q\big)^{-1}\mathcal{Q}(X)^{-1}\delta S^{\top}\\
&\quad - F^{\top}\delta Q\big(I+\mathcal{Q}(X)^{-1}\delta Q\big)^{-1}\mathcal{Q}(X)^{-1}\delta S^{\top}\\
&\quad - \delta S\mathcal{Q}(X)^{-1}\delta Q\big(I+\mathcal{Q}(X)^{-1}\delta Q\big)^{-1}F\\
&\quad + F^{\top}\delta QF - F^{\top}\delta Q\big(I+\mathcal{Q}(X)^{-1}\delta Q\big)^{-1}\mathcal{Q}(X)^{-1}\delta QF.
\end{aligned}$$

Upon substituting (10) into δSF and $F^{\top}\delta QF$, we have

$$\begin{aligned}
\delta SF &= \Delta SF + C^{\top}X\Delta DF + \Delta C^{\top}XDF + \Delta C^{\top}X\Delta DF + X\Delta BF,\\
F^{\top}\delta QF &= F^{\top}\Delta RF + F^{\top}D^{\top}X\Delta DF + F^{\top}\Delta D^{\top}XDF + F^{\top}\Delta D^{\top}X\Delta DF,
\end{aligned}$$

so that the structure of E in (17) can be partitioned into linear equations

$$\begin{aligned}
E_1 &:= -\big(K^{\top}\Delta DF + F^{\top}\Delta D^{\top}K + K^{\top}\Delta C + \Delta C^{\top}K + F^{\top}\Delta RF + F^{\top}\Delta S^{\top} + \Delta SF + \Delta H\big),\\
E_2 &:= -\big[\Delta A^{\top}X + X\Delta A + \Delta C^{\top}X\Delta C + F^{\top}\Delta B^{\top}X + X\Delta BF\\
&\quad + F^{\top}\Delta D^{\top}X\Delta C + \Delta C^{\top}X\Delta DF + F^{\top}\Delta D^{\top}X\Delta DF\\
&\quad - \big(F^{\top}\delta Q + \delta S\big)\big(I+\mathcal{Q}(X)^{-1}\delta R\big)^{-1}\mathcal{Q}(X)^{-1}\big(\delta QF + \delta S^{\top}\big)\big],
\end{aligned}$$

that is, $E = E_1 + E_2$.

Lemma 2.1 *Let X be the stabilizing solution of SARE* (1a)-(1b) *and $\widetilde{X}$ be a symmetric solution of perturbed SARE* (5a)-(5b). *If $\Delta X = \widetilde{X} - X$, then ΔX satisfies the equation*

$$\Phi^{\top}\Delta X + \Delta X\Phi + \Psi^{\top}\Delta X\Psi = E_1 + E_2 + h_1(\Delta X) + h_2(\Delta X), \tag{20}$$

where

$$\begin{aligned}
E_1 &= -\big(K^{\top}\Delta DF + F^{\top}\Delta D^{\top}K + K^{\top}\Delta C + \Delta C^{\top}K + F^{\top}\Delta RF\\
&\quad + F^{\top}\Delta S^{\top} + \Delta SF + \Delta H\big), &(21a)\\
E_2 &= -\big[\Delta A^{\top}X + X\Delta A + \Delta C^{\top}X\Delta C + F^{\top}\Delta B^{\top}X + X\Delta BF\\
&\quad + F^{\top}\Delta D^{\top}X\Delta C + \Delta C^{\top}X\Delta DF + F^{\top}\Delta D^{\top}X\Delta DF\\
&\quad - \big(F^{\top}\delta Q + \delta S\big)\big(I+\mathcal{Q}(X)^{-1}\delta R\big)^{-1}\mathcal{Q}(X)^{-1}\big(\delta QF + \delta S^{\top}\big)\big],\\
h_1(\Delta X) &= -\big(\Delta\Phi^{\top}\Delta X + \Delta X\Delta\Phi + \Psi^{\top}\Delta X\Delta\Psi + \Delta\Psi^{\top}\Delta X\Psi + \Delta\Psi^{\top}\Delta X\Delta\Psi\big), &(21b)\\
h_2(\Delta X) &= \widetilde{\Psi}^{\top}\Delta X\widetilde{D}\Omega\widetilde{D}^{\top}\Delta X\widetilde{\Psi} + \Delta X\widetilde{B}\Omega\widetilde{B}^{\top}\Delta X\\
&\quad + \Delta X\widetilde{B}\Omega\widetilde{D}^{\top}\Delta X\widetilde{\Psi} + \widetilde{\Psi}^{\top}\Delta X\widetilde{D}\Omega\widetilde{B}^{\top}\Delta X, &(21c)
\end{aligned}$$

where $\Omega = (I + \widetilde{\mathcal{Q}}(X)^{-1}\widetilde{D}^{\top}\Delta X\widetilde{D})^{-1}\widetilde{\mathcal{Q}}(X)^{-1}$, *the matrices* ΔA, ΔB, *and so on are given by* (9)-(12).

Note that E_1 and E_2 are not dependent on ΔX, $h_1(\Delta X)$ is a linear function of ΔX, and $h_2(\Delta X)$ is a function of ΔX with degree at most 2. Assume that the linear operator $\mathcal{L}_c$ of (4) is invertible. It is easy to see that the perturbed equation (20) is true if and only if

$$\Delta X = \mathcal{L}_c^{-1}E_1 + \mathcal{L}_c^{-1}E_2 + \mathcal{L}_c^{-1}h_1(\Delta X) + \mathcal{L}_c^{-1}h_2(\Delta X). \tag{22}$$

Thus far, we have not specified the condition for the existence of the solution ΔX in (22). In the subsequent discussion, we shall limit our attention to identifying the condition of the existence of a fixed point of (23), that is, to determine an upper bound on the size of ΔX.

3 Perturbation bounds

Let $f : \mathcal{S}^n \to \mathcal{S}^n$ be a continuous mapping defined by

$$f(Y) = \mathcal{L}_c^{-1}E_1 + \mathcal{L}_c^{-1}E_2 + \mathcal{L}_c^{-1}h_1(Y) + \mathcal{L}_c^{-1}h_2(Y) \quad \text{for } Y \in \mathcal{S}^n. \tag{23}$$

We see that any fixed point of the mapping f is a solution to the perturbed equation (22). Our approach in this section is to present an upper bound for the existence of some fixed points ΔX. It starts with the discussion that the mapping f given by (23) satisfies

$$\|f(\Delta X)\|_F \le \|\mathcal{L}_c^{-1}E_1\|_F + \|\mathcal{L}_c^{-1}E_2\|_F + \|\mathcal{L}_c^{-1}h_1(\Delta X)\|_F + \|\mathcal{L}_c^{-1}h_2(\Delta X)\|_F.$$

Define linear operators $\mathcal{M} : \mathbb{R}^{n\times n} \to \mathcal{S}^n, \mathcal{N} : \mathbb{R}^{n\times m} \to \mathcal{S}^n, \mathcal{T} : \mathcal{S}^m \to \mathcal{S}^n$ and $\mathcal{H} : \mathbb{R}^{n\times m} \to \mathcal{S}^n$ by

$$\mathcal{M}\Delta C = \mathcal{L}_c^{-1}\big(K^{\top}\Delta C + \Delta C^{\top}K\big), \tag{24a}$$

$$\mathcal{N}\Delta D = \mathcal{L}_c^{-1}\big(K^{\top}\Delta DF + F^{\top}\Delta D^{\top}K\big), \tag{24b}$$

$$\mathcal{T}\Delta R = \mathcal{L}_c^{-1}\big(F^{\top}\Delta RF\big), \tag{24c}$$

$$\mathcal{H}\Delta S = \mathcal{L}_c^{-1}\big(F^{\top}\Delta S^{\top} + \Delta SF\big), \tag{24d}$$

and the scalars $\omega, \mu, \nu, \tau, \eta$ by

$$\omega = \|\mathcal{L}_c^{-1}\|, \qquad \mu = \|\mathcal{M}\|, \qquad \nu = \|\mathcal{N}\|, \qquad \tau = \|\mathcal{T}\|, \qquad \eta = \|\mathcal{H}\|. \tag{25}$$

From (21a) we then have

$$\|\mathcal{L}_c^{-1}E_1\|_F \le \mu\|\Delta C\|_F + \nu\|\Delta D\|_F + \tau\|\Delta S\|_F + \eta\|\Delta R\|_F + \omega\|\Delta H\| \equiv \varepsilon_1. \tag{26}$$

We now move into more specific details pertaining to the discussion of the fixed point of the continuous mapping f. Before doing so, we need to describe an important property of the norm of the product of two matrices and repeatedly employ it in the following discussion. For the proof, the reader is referred to [30, Theorem 3.9].

Lemma 3.1 *Let A and B be two matrices in $\mathbb{R}^{n\times n}$. Then $\|AB\|_F \le \|A\|_2\|B\|_F$ and $\|AB\|_F \le \|A\|_F\|B\|_2$.*

It immediately follows that the matrices δQ and δS, defined by (10), satisfy

$$\begin{aligned}
&\|\delta Q\|_F \le \|\Delta R\|_F + 2\|XD\|_2\|\Delta D\|_F + \|X\|_2\|\Delta D\|_F^2 \equiv \delta_r,\\
&\|\delta S\|_F \le \|\Delta S\|_F + \|XC\|_2\|\Delta D\|_F + \|XD\|_2\|\Delta C\|_F\\
&\qquad + \|X\|_2\|\Delta D\|_F\|\Delta C\|_F + \|X\|_2\|\Delta B\|_F \equiv \delta_s.
\end{aligned} \tag{27}$$

Assume that the scalar δ_r satisfies

$$1 - \left\|\mathcal{Q}(X)^{-1}\right\|_2\delta_r > 0. \tag{28}$$

Then $\|\mathcal{L}_c^{-1}E_2\|_F$ is bounded by

$$\begin{aligned}
\left\|\mathcal{L}_c^{-1}E_2\right\|_F &\le 2\omega\|X\|_2\big(\|\Delta A\|_F + \|F\|_2\|\Delta B\|_F\big) + \omega\|X\|_2\big(\|\Delta C\|_F + \|F\|_2\|\Delta D\|_F\big)^2\\
&\quad + \frac{\omega\|\mathcal{Q}(X)^{-1}\|_2(\|F\|_2\delta_r + \delta_s)^2}{1 - \|\mathcal{Q}(X)^{-1}\|_2\delta_r} \equiv \varepsilon_2.
\end{aligned} \tag{29}$$

From (21b) we see that

$$\left\|h_1(\Delta X)\right\|_F \le \big(2\|\Delta\Phi\|_F + 2\|\Psi\|_2\|\Delta\Psi\|_F + \|\Delta\Psi\|_F^2\big)\|\Delta X\|_F,$$

and also from (18) and (19) we have

$$\begin{aligned}
\|\Delta\Phi\|_F &\le \|\Delta A\|_F + \|F\|_2\|\Delta B\|_F + \big(\left\|B\mathcal{Q}(X)^{-1}\right\|_2 + \left\|\mathcal{Q}(X)^{-1}\right\|_2\|\Delta B\|_F\big)\delta_s\\
&\quad + \frac{\|\mathcal{Q}(X)^{-1}\|_2(\|B\|_2 + \|\Delta B\|_F)(\|F\|_2 + \|\mathcal{Q}(X)^{-1}\|_2\delta_s)\delta_r}{1 - \|\mathcal{Q}(X)^{-1}\|_2\delta_r} \equiv \delta_\Phi,
\end{aligned} \tag{30}$$

$$\begin{aligned}
\|\Delta\Psi\|_F &\le \|\Delta C\|_F + \|F\|_2\|\Delta D\|_F + \big(\left\|D\mathcal{Q}(X)^{-1}\right\|_2 + \left\|\mathcal{Q}(X)^{-1}\right\|_2\|\Delta D\|_F\big)\delta_s\\
&\quad + \frac{\|\mathcal{Q}(X)^{-1}\|_2(\|D\|_2 + \|\Delta D\|_F)(\|F\|_2 + \|\mathcal{Q}(X)^{-1}\|_2\delta_s)\delta_r}{1 - \|\mathcal{Q}(X)^{-1}\|_2\delta_r} \equiv \delta_\Psi.
\end{aligned} \tag{31}$$

It follows that

$$\left\|\mathcal{L}_c^{-1}h_1(\Delta X)\right\|_F \le \omega\delta\|\Delta X\|_F, \tag{32}$$

where the positive scalar δ is defined by

$$\delta = 2\delta_\Phi + 2\psi\delta_\Psi + \delta_\Psi^2 \quad \text{with } \|\Psi\|_2 = \psi. \tag{33}$$

Also, from (28) and Lemma 3.1 we know that $\widetilde{\mathcal{Q}}(X) = \mathcal{Q}(X) + \delta Q = \mathcal{Q}(X)^{-1}(I + \mathcal{Q}(X)^{-1}\delta Q)$ and $\|\mathcal{Q}(X)^{-1}\delta Q\|_F \le \|\mathcal{Q}(X)^{-1}\|_2\delta_r < 1$. This implies that $\widetilde{\mathcal{Q}}(X)$ is nonsingular,

$$\begin{aligned}
\|\tilde{B}\|_2^2\left\|\widetilde{\mathcal{Q}}(X)^{-1}\right\|_2 &= \|B + \Delta B\|_2^2\left\|\big(\mathcal{Q}(X) + \delta Q\big)^{-1}\right\|_2\\
&= \|B + \Delta B\|_2^2\left\|\big(I + \mathcal{Q}(X)^{-1}\delta Q\big)^{-1}\mathcal{Q}(X)^{-1}\right\|_2\\
&\le \frac{\|\mathcal{Q}(X)^{-1}\|_2(\|B\|_2 + \|\Delta B\|_F)^2}{1 - \|\mathcal{Q}(X)^{-1}\|_2\delta_r} \equiv \gamma_B.
\end{aligned} \tag{34}$$

Similarly, we have

$$\|\tilde{D}\|_2^2 \left\| \widetilde{\mathcal{Q}}(X)^{-1} \right\|_2 \leq \frac{\|\mathcal{Q}(X)^{-1}\|_2 (\|D\|_2 + \|\Delta D\|_F)^2}{1 - \|\mathcal{Q}(X)^{-1}\|_2 \delta_r} \equiv \gamma_D. \tag{35}$$

Assume that

$$1 - \gamma_D \|\Delta X\|_F > 0. \tag{36}$$

It then follows from Lemma 3.1 and (21c) that

$$\left\| h_2(\Delta X) \right\|_F \leq \frac{(\|\widetilde{\Psi}\|_2 \|\widetilde{D}\|_2 + \|\widetilde{B}\|_2)^2 \|\widetilde{\mathcal{Q}}(X)^{-1}\|_2 \|\Delta X\|_F^2}{1 - \|\widetilde{D}\|_2^2 \|\widetilde{\mathcal{Q}}(X)^{-1}\|_2 \|\Delta X\|_F}, \tag{37}$$

and from (9), (11) and (31) that

$$\begin{aligned} &\|\widetilde{B}\|_2 \leq \|B\|_2 + \|\Delta B\|_2 \equiv \alpha_B, \\ &\|\widetilde{D}\|_2 \leq \|D\|_2 + \|\Delta D\|_2 \equiv \alpha_D, \\ &\|\widetilde{\Psi}\|_2 \leq \|\Psi\|_2 + \|\Delta \Psi\|_F \leq \|\Psi\|_2 + \delta_\Psi \equiv \tilde{\psi}. \end{aligned} \tag{38}$$

Upon substituting (34), (35) and (38) into (37), we see that

$$\left\| \mathcal{L}_c^{-1} h_2(\Delta X) \right\|_F \leq \frac{\omega(\tilde{\psi}^2 \gamma_D + 2\tilde{\psi} \alpha_B \alpha_D + \gamma_B) \|\Delta X\|_F^2}{1 - \gamma_D \|\Delta X\|_F}.$$

Finally, by (26), (29) and (32), we arrive at the statement

$$\left\| f(\Delta X) \right\|_F \leq \varepsilon + \omega\delta \|\Delta X\|_F + \frac{\omega\alpha \|\Delta X\|_F^2}{1 - \gamma_D \|\Delta X\|_F}, \tag{39}$$

where

$$\alpha \equiv \tilde{\psi}^2 \gamma_D + 2\tilde{\psi} \alpha_B \alpha_D + \gamma_B, \qquad \varepsilon \equiv \varepsilon_1 + \varepsilon_2. \tag{40}$$

Consider the quadratic equation

$$(\gamma_D - \omega\delta\gamma_D + \omega\alpha)\xi^2 - (1 - \omega\delta + \varepsilon\gamma_D)\xi + \varepsilon = 0. \tag{41}$$

It is true that if

$$\delta < \frac{1}{\omega}, \tag{42a}$$

$$\varepsilon \leq \frac{(1 - \omega\delta)^2}{\gamma_D - \omega\delta\gamma_D + 2\omega\alpha + \sqrt{(\gamma_D - \omega\delta\gamma_D + 2\omega\alpha)^2 - \gamma_D^2 (1 - \omega\delta)^2}}, \tag{42b}$$

then the positive scalar ξ_* denoted by

$$\xi_* = \frac{2\varepsilon}{(1 - \omega\delta + \varepsilon\gamma_D) + \sqrt{(1 - \omega\delta + \varepsilon\gamma_D)^2 - 4(\gamma_D - \omega\delta\gamma_D + \omega\alpha)\varepsilon}} \tag{43}$$

is a solution to (41). Let $\mathcal{S}^n_{\xi_*}$ be a compact subset of $\mathcal{S}^n$ given by

$$\mathcal{S}^n_{\xi_*} = \{\Delta X \in \mathcal{S}^n : \|\Delta X\|_F \le \xi_*\}.$$

It can be seen that in (39)

$$\|f(\Delta X)\|_F \le \xi_* \quad \text{if } \Delta X \in \mathcal{S}^n_{\xi_*}.$$

It then follows from the Brouwer fixed-point theorem (see [31]) that the continuous mapping f has a fixed point $\Delta X_* \in \mathcal{S}^n_{\xi_*}$, that is, condition (22) automatically holds.

Observe also that if $\Delta X \in \mathcal{S}^n_{\xi_*}$, then

$$\begin{aligned}
1 - \gamma_D\|\Delta X\|_F &\ge (1 - \gamma_D\xi_*) \quad \text{(by (43))} \\
&\ge 1 - \frac{2\varepsilon\gamma_D}{1 - \omega\delta + \varepsilon\gamma_D} = \frac{1 - \omega\delta - \varepsilon\gamma_D}{1 - \omega\delta + \varepsilon\gamma_D} \quad \text{(by (42b))} \\
&\ge \frac{1 - \omega\delta - \frac{(1-\omega\delta)^2\gamma_D}{\gamma_D - \omega\delta\gamma_D + 2\omega\alpha}}{1 - \omega\delta + \varepsilon\gamma_D} \quad \text{(by (42a))} \\
&= \frac{2(1 - \omega\delta)\omega\alpha}{(1 - \omega\delta + \varepsilon\gamma_D)(\gamma_D - \omega\delta\gamma_D + 2\omega\alpha)} \ge 0.
\end{aligned}$$

This implies that assumption (36) is true, if assumption (42a)-(42b) is true.

4 Stability analysis

We have shown that the mapping f given by (23) has a Hermitian fixed point ΔX_*. This further implies that perturbed SARE (1a)-(1b) has a Hermitian solution $\widetilde{X} = X + \Delta X_*$. In this section, we want to discuss the stability of the solution $\widetilde{X}$, *i.e.*, show that the solution $\widetilde{X}$ is the unique maximal solution to SARE (1a)-(1b). Let Υ and Π be two operators defined by

$$\Upsilon(W) = \Phi^\top W + W\Phi, \qquad \Pi(W) = \Psi^\top W\Psi, \quad W \in \mathcal{S}^n,$$

with the notations Φ and Ψ given in Definition 1.1. It follows that the operator $\mathcal{L}_c$ defined by (4) can also be written as

$$\mathcal{L}_c(W) = \Upsilon(W) + \Pi(W), \quad W \in \mathcal{S}^n. \tag{44}$$

We then have the following important result addressing the condition for a linear operator to be stable. To see a few necessary and sufficient conditions on the stability, we refer to the results and proofs given in [3].

Theorem 4.1 *The linear operator $\mathcal{L}_c = \Upsilon + \Pi$ given by* (44) *is stable, i.e., $\sigma(\mathcal{L}_c) \subset \mathbb{C}_-$, if and only if $\sigma(\Phi) \subset \mathbb{C}_-$ and $\det(\Upsilon + \tau\Pi) \ne 0$ for all $\tau \in [0,1]$.*

When small perturbations $Z_1, Z_2 \in \mathbb{R}^{n\times n}$ are taken into consideration, the perturbed operator of $\mathcal{L}_c$ can be expressed by

$$\widetilde{\mathcal{L}}_c(W) = \widetilde{\Upsilon}(W) + \widetilde{\Pi}(W), \tag{45}$$

where $\widetilde{\Upsilon}(W) = (\Phi + Z_1)^\top W + W(\Phi + Z_1)$ and $\widetilde{\Pi}(W) = (\Psi + Z_2)^\top W(\Psi + Z_2)$ for all $W \in \mathcal{S}^n$. Define the quantity

$$\ell(\theta) = \left\|(\Upsilon + \theta\Pi)^{-1}\right\|$$

for $\theta \in [0,1]$ and

$$\beta(\mathcal{L}_c) = \min_{(Z_1,Z_2)\in\mathcal{Z}} \max\left\{\|Z_1\|, \|Z_2\|\right\},$$

where the set $\mathcal{Z} = \{(Z_1, Z_2) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \mid \det(\widetilde{\Upsilon} + \theta\widetilde{\Pi}) = 0 \text{ for some } \theta \in [0,1]\}$. It should be noted that if $\sigma(\Phi) \subset \mathbb{C}_-$, $\Psi = 0$ and $Z_2 = 0$, then

$$\beta(\mathcal{L}_c) \le \beta(\Phi), \tag{46}$$

where the value $\beta(\Phi)$ is defined by [27]

$$\beta(\Phi) = \min\left\{\|Z_1\| \;\middle|\; \max_{1\le j\le n} \operatorname{Re}\lambda_j(\Phi + Z_1) = 0, Z_1 \in \mathbb{R}^{n\times n}\right\}. \tag{47}$$

Here, $\lambda_j(\Phi + Z_1)$ $(j = 1, \ldots, n)$ denote the eigenvalues of $\Phi + Z_1$.

The connection between $\beta(\mathcal{L}_c)$ and the maximum of the scalar function $\ell(\theta)$ on $[0,1]$ can be established in the following form.

Theorem 4.2 [23] *Suppose that the linear operator $\mathcal{L}_c$ given by* (44) *is stable, and let*

$$\ell_c = \max_{\theta\in[0,1]} \ell(\theta), \qquad \psi = \|\Psi\|. \tag{48}$$

Then

$$\beta(\mathcal{L}_c) \ge \frac{\ell_c^{-1}}{(\psi + 1) + \sqrt{(\psi + 1)^2 + \ell_c^{-1}}}.$$

We now apply Theorem 4.2 to (46) and obtain that

$$\beta(\Phi) \ge \beta(\mathcal{L}_c) \ge \frac{\ell_c^{-1}}{(\psi + 1) + \sqrt{(\psi + 1)^2 + \ell_c^{-1}}}.$$

Hence, if a perturbation matrix $Z_1 \in \mathbb{R}^{n\times n}$ satisfies

$$\|Z_1\| < \frac{\ell_c^{-1}}{(\psi + 1) + \sqrt{(\psi + 1)^2 + \ell_c^{-1}}},$$

then (47) implies that the matrix $\Phi + Z_1$ must be c-stable.

We now turn to a key stability test of the operator $\mathcal{L}_c$, the striking tool of our stability analysis.

Theorem 4.3 [23] *Suppose that the linear operator $\mathcal{L}_c$ is stable, and let the scalars ℓ_c and ψ be defined as in* (48). *If the perturbation matrices $Z_1, Z_2 \in \mathbb{R}^{n\times n}$ satisfy*

$$\max\left\{\|Z_1\|, \|Z_2\|\right\} < \frac{\ell_c^{-1}}{(\psi + 1) + \sqrt{(\psi + 1)^2 + \ell_c^{-1}}}, \tag{49}$$

then $\Phi + Z_1$ *is c-stable and the perturbed linear operator* $\widetilde{\mathcal{L}}_c$ *defined by* (45) *is also stable, i.e.,* $\sigma(\widetilde{\mathcal{L}}_c) \subset \mathbb{C}_-$.

Upon substituting $\widetilde{X}$ for X in $\widetilde{\mathcal{S}}(X)$ and $\widetilde{\mathcal{Q}}(X)$ of (11), we shall have

$$\widetilde{\mathcal{Q}}(\widetilde{X}) = \widetilde{R} + \widetilde{D}^\top \widetilde{X}\widetilde{D} = \mathcal{Q}(X) + \delta Q + \widetilde{D}^\top \Delta X \widetilde{D} \equiv \mathcal{Q}(X) + \Delta R, \tag{50}$$

$$\begin{aligned}\widetilde{\mathcal{S}}(\widetilde{X}) &= \widetilde{S} + \widetilde{C}^\top \widetilde{X}\widetilde{D} + \widetilde{X}\widetilde{B} = \mathcal{S}(X) + \delta S + \Delta X \widetilde{B} + \widetilde{C}^\top \Delta X \widetilde{D} \\ &\equiv \mathcal{S}(X) + \Delta S.\end{aligned} \tag{51}$$

Also, corresponding to $\widetilde{X}$, the perturbed $\Psi_{\widetilde{X}}$ and $\Phi_{\widetilde{X}}$ of Ψ and Φ, respectively, can be expressed in terms of the formulae

$$\begin{aligned}\Phi_{\widetilde{X}} &= (\Delta A + A) - (\Delta B + B)(\Gamma + \Delta S)^{-1}(\Lambda + \Delta R)^\top := \Phi + \Delta\Phi, \\ \Psi_{\widetilde{X}} &= (\Delta C + C) - (\Delta D + D)(\Gamma + \Delta R)^{-1}(\Lambda + \Delta S)^\top := \Psi + \Delta\Psi,\end{aligned} \tag{52}$$

with

$$\begin{aligned}\Delta\Phi &:= \Delta A - \Delta B\mathcal{Q}(X)^{-1}\mathcal{S}(X)^\top - \Delta B\mathcal{Q}(X)^{-1}\Delta S^\top - B\mathcal{Q}(X)^{-1}\Delta S^\top \\ &\quad + (\Delta B + B)\mathcal{Q}(X)^{-1}\Delta R\mathcal{Q}(X)^{-1}\big(I + \mathcal{Q}(X)^{-1}\Delta R\big)^{-1}\big(\mathcal{S}(X)^\top + \Delta S^\top\big), \\ \Delta\Psi &:= \Delta C - \Delta D\mathcal{Q}(X)^{-1}\mathcal{S}(X)^\top - \Delta D\mathcal{Q}(X)^{-1}\Delta S^\top - D\mathcal{Q}(X)^{-1}\Delta S^\top \\ &\quad + (\Delta D + D)\mathcal{Q}(X)^{-1}\Delta R\mathcal{Q}(X)^{-1}\big(I + \mathcal{Q}(X)^{-1}\Delta R\big)^{-1}\big(\mathcal{S}(X)^\top + \Delta S^\top\big).\end{aligned}$$

Let $\alpha_C := \|C\|_2 + \|\Delta C\|_2$. Since $\|\Delta X\|_F \le \xi_*$, it follows from (38), (50) and (51) that

$$\begin{aligned}&\|\Delta R\|_F \le \delta_r + \alpha_D^2 \xi_* := c_r, \\ &\|\Delta S\|_F \le \delta_s + (\alpha_B + \alpha_C\alpha_D)\xi_* := c_s.\end{aligned}$$

Thus $\|\Delta\Phi\|_F$ and $\|\Delta\Psi\|_F$ are bounded by the inequalities

$$\begin{aligned}\|\Delta\Phi\|_F &\le \|\Delta A\|_F + \|F\|_2\|\Delta B\|_F + \frac{\alpha_B\|\mathcal{Q}(X)^{-1}\|_2(c_s + c_r\|F\|_2)}{1 - c_r\|\mathcal{Q}(X)^{-1}\|_2}, \\ \|\Delta\Psi\|_F &\le \|\Delta C\|_F + \|F\|_2\|\Delta D\|_F + \frac{\alpha_D\|\mathcal{Q}(X)^{-1}\|_2(c_s + c_r\|F\|_2)}{1 - c_r\|\mathcal{Q}(X)^{-1}\|_2}.\end{aligned}$$

Here, the above upper bounds are obtained by simplifying those given by (30) and (31). Let

$$\begin{aligned}&f = \|F\|_2, \qquad \gamma = \big\|\mathcal{Q}(X)^{-1}\big\|_2, \\ &\alpha_C = \|C\|_2 + \|\Delta C\|_F, \qquad \zeta_1 = \max\big\{\|\Delta A\|_F, \|\Delta C\|_F\big\}, \\ &\zeta_2 = \max\big\{\|\Delta B\|_F, \|\Delta D\|_F\big\}, \qquad \zeta_3 = \max\{\alpha_B, \alpha_D\},\end{aligned} \tag{53}$$

where α_B and α_D are defined by (38) and Θ is defined to be the right-hand side of (49), that is,

$$\Theta = \frac{\ell_c^{-1}}{(\psi + 1) + \sqrt{(\psi + 1)^2 + \ell_c^{-1}}}. \tag{54}$$

We then have

$$\max\{\|\Delta\Phi\|_F, \|\Delta\Psi\|_F\} \le \zeta_1 + f\zeta_2 + \frac{\gamma\zeta_3(\delta_s + \delta_r f) + \gamma\zeta_3(\alpha_B + \alpha_C\alpha_D + \alpha_D^2 f)\xi_*}{1 - \delta_r f - \alpha_D^2 f\xi_*}.$$

It follows that if the condition

$$\zeta_1 + f\zeta_2 + \frac{\gamma\zeta_3(\delta_s + \delta_r f) + \gamma\zeta_3(\alpha_B + \alpha_C\alpha_D + \alpha_D^2 f)\xi_*}{1 - \delta_r f - \alpha_D^2 f\xi_*} < \Theta$$

or, equivalently,

$$\xi_* < \frac{(\Theta - \zeta_1 - f\zeta_2)(1 - \delta_r f) - \gamma\zeta_3(\delta_s + \delta_r f)}{(\Theta - \zeta_1 - f\zeta_2)\alpha_D^2 f + \gamma\zeta_3(\alpha_B + \alpha_C\alpha_D + \alpha_D^2 f)}$$

holds, then corresponding to Theorem 4.3, the perturbed linear operator $\widetilde{\mathcal{L}}_c$ with respect to $\widetilde{X}$ is stable. In other words, the matrix $\widetilde{X} \in \mathcal{S}^n$ must be the unique stabilizing (and maximal) solution to perturbed SARE (5a)-(5b).

We now have all the materials needed for the existence of a stabilizing solution of (5a)-(5b).

Theorem 4.4 (Perturbation bound) *Let X be the stabilizing solution of* (1a)-(1b). *Let ω, δ_r, δ_s, δ, γ_D, α_B, α_D, α, ε, f, γ, α_C, ζ_1, ζ_2, ζ_3, Θ be the scalars defined by* (25), (27), (33), (35), (38), (40), (53) *and* (54), *respectively. Define*

$$\xi_* = \frac{2\varepsilon}{(1 - \omega\delta + \varepsilon\gamma_D) + \sqrt{(1 - \omega\delta + \varepsilon\gamma_D)^2 - 4(\gamma_D - \omega\delta\gamma_D + \omega\alpha)\varepsilon}}.$$

If the perturbed quantities of the coefficients of (5a)-(5b) *are sufficiently small, for example, $\varepsilon \ll 1$, such that*

$$1 - \left\|\mathcal{Q}(X)^{-1}\right\|_2 \delta_r > 0,$$

$$1 - \omega\delta > 0,$$

$$\frac{(1 - \omega\delta)^2}{\gamma_D - \omega\delta\gamma_D + 2\omega\alpha + \sqrt{(\gamma_D - \omega\delta\gamma_D + 2\omega\alpha)^2 - \gamma_D^2(1 - \omega\delta)^2}} - \varepsilon \ge 0,$$

$$\frac{(\Theta - \zeta_1 - f\zeta_2)(1 - \delta_r f) - \gamma\zeta_3(\delta_s + \delta_r f)}{(\Theta - \zeta_1 - f\zeta_2)\alpha_D^2 f + \gamma\zeta_3(\alpha_B + \alpha_C\alpha_D + \alpha_D^2 f)} - \xi_* > 0,$$

then perturbed SARE (5a)-(5b) *has the unique stabilizing solution $\widetilde{X}$, and*

$$\frac{\|\widetilde{X} - X\|_F}{\|X\|_F} \le \frac{\xi_*}{\|X\|_F}. \tag{55}$$

5 Condition number of the SARE

In the study of a computational problem, a fundamental issue is to determine the condition number of a problem to be the ratio of the relative change in the solution to the relative

change in the argument. Applying the theory of condition number given by Rice [32], we define the condition number $c(X)$ of the stabilizing solution X of SARE (1a)-(1b) by

$$c(X) = \lim_{\delta \to 0^+} \sup_{\Omega_\delta} \frac{\|\Delta X\|_F}{\kappa \delta}, \tag{56}$$

where the set of perturbed matrices Ω_δ is defined by

$$\begin{aligned} \Omega_\delta &= \Omega_\delta(\kappa_A, \kappa_B, \kappa_C, \kappa_D, \kappa_S, \kappa_R, \kappa_H) \\ &= \big\{(\Delta A, \Delta B, \Delta C, \Delta D, \Delta S) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times m}, \\ &\quad (\Delta R, \Delta H) \in \mathcal{S}^n \times \mathcal{S}^m \mid 0 < \delta_p \le \delta\big\}, \end{aligned} \tag{57}$$

with

$$\delta_p = \left\| \left(\frac{\Delta A}{\kappa_A}, \frac{\Delta B}{\kappa_B}, \frac{\Delta C}{\kappa_C}, \frac{\Delta D}{\kappa_D}, \frac{\Delta S}{\kappa_S}, \frac{\Delta R}{\kappa_R}, \frac{\Delta H}{\kappa_H} \right) \right\|_F,$$

and κ_A, κ_B, κ_C, κ_D, κ_S, κ_R, κ_H, κ are positive parameters. Then (56) gives the absolute condition number $c_{\mathrm{abs}}(X)$ if

$$(\kappa_A, \kappa_B, \kappa_C, \kappa_D, \kappa_S, \kappa_R, \kappa_H, \kappa) = (1,1,1,1,1,1,1,1)$$

and gives the relative condition number $c_{\mathrm{rel}}(X)$ if

$$(\kappa_A, \kappa_B, \kappa_C, \kappa_D, \kappa_S, \kappa_R, \kappa_H, \kappa) = \big(\|A\|_F, \|B\|_F, \|C\|_F, \|D\|_F, \|S\|_F, \|R\|_F, \|H\|_F, \|X\|_F\big).$$

It follows from (22) and (24a)-(24d) that

$$\Delta X = \mathcal{P}\Delta A + \mathcal{Q}\Delta B + \mathcal{M}\Delta C + \mathcal{N}\Delta D + \mathcal{H}\Delta S + \mathcal{T}\Delta R + \mathcal{L}_c^{-1}\Delta H + O\big(\delta_p^2\big),$$

where the linear operators $\mathcal{P} : \mathbb{R}^{n\times n} \to \mathcal{S}^n$ and $\mathcal{Q} : \mathbb{R}^{n\times m} \to \mathcal{S}^n$ are defined by

$$\mathcal{P}\Delta A = \mathcal{L}_c^{-1}\big(X\Delta A + \Delta A^\top X\big), \tag{58a}$$

$$\mathcal{Q}\Delta B = \mathcal{L}_c^{-1}\big(X\Delta B F + F^\top \Delta B^\top X\big). \tag{58b}$$

In order to derive the explicit expression for the condition number $c(X)$ of the stabilizing solution X of (1a)-(1b), we require a theorem concerning the form of the optimal solution. This theorem can be regarded as a theoretical extension of the results discussed in [25, 33]. Most strategies have been established earlier by using much heavier machinery. Since this theorem is most relevant to our stability analysis, we briefly outline a direct proof with ideas from [34] to make this presentation more self-contained.

Theorem 5.1 *Let $\mathcal{L} : \mathbb{R}^{n\times n} \times \mathbb{R}^{m\times m} \to \mathbb{R}^{k\times k}$ be a linear operator and*

$$\mathcal{L}(Z_1, Z_2)^\top = \mathcal{L}\big(Z_1, Z_2^\top\big) \tag{59}$$

for all $Z_1 \in \mathbb{R}^{n\times n}$ *and* $Z_2 \in \mathbb{R}^{m\times m}$. *Then the optimal solution* $(Z_{1\star}, Z_{2\star})$ *to the problem*

$$\max_{\|(Z_1,Z_2)\|_F=1} \left\| \mathcal{L}(Z_1, Z_2) \right\|_F \tag{60}$$

exists for some $Z_{1\star} \in \mathbb{R}^{n\times n}$ *and* $Z_{2\star} = \pm Z_{2\star}^\top \in \mathbb{R}^{m\times m}$. *Furthermore, if the linear operator* $\mathcal{L}(0, Z_2)$ *is a positive operator with respect to any* $Z_2 \in \mathbb{C}^{m\times m}$, *that is, for all* $Z_2 \in \mathbb{C}^{m\times m}$, *we have*

$$\mathcal{L}(0, Z_2) \succeq 0 \quad \textit{if } Z_2 \succeq 0.$$

Then there exists an optimal solution $(Z_{1\star}, Z_{2\star}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{m\times m}$ *to problem* (60) *such that* $Z_{2\star}$ *is symmetric*.

Proof Since $\mathcal{L}$ is a linear operator on a finite dimensional space, it is clear that the optimal solution of (60) exists. Assume that $(Z_{1\star}, Z_{2\star})$ solves this optimization problem. Let $\sigma_{\max} = \|L(Z_{1\star}, Z_{2\star})\|_F$ and $L \in \mathbb{R}^{k^2\times(n^2+m^2)}$ be the matrix representation of the operator $\mathcal{L}$ such that

$$\operatorname{vec}\big(\mathcal{L}(Z_1, Z_2)\big) = L \begin{bmatrix} \operatorname{vec}(Z_1) \\ \operatorname{vec}(Z_2) \end{bmatrix}. \tag{61}$$

By (59) and (61), we have

$$\begin{bmatrix} \operatorname{vec}(Z_{1\star}) \\ \operatorname{vec}(Z_{2\star}) \end{bmatrix}^\top L^\top L \begin{bmatrix} \operatorname{vec}(Z_{1\star}) \\ \operatorname{vec}(Z_{2\star}) \end{bmatrix} = \begin{bmatrix} \operatorname{vec}(Z_{1\star}) \\ \operatorname{vec}(Z_{2\star}^\top) \end{bmatrix}^\top L^\top L \begin{bmatrix} \operatorname{vec}(Z_{1\star}) \\ \operatorname{vec}(Z_{2\star}^\top) \end{bmatrix} = \sigma_{\max}^2. \tag{62}$$

Note that

$$\ell := \|Z_{1\star}\|_F^2 + \left\| \frac{1}{2}\big(Z_{2\star} + Z_{2\star}^\top\big) \right\|_F^2 = 0, \quad \text{only if } Z_{1\star} = 0, Z_{2\star} = -Z_{2\star}^\top.$$

It follows that if $Z_{2\star}^\top \neq -Z_{2\star}$, by (62), we see that $\frac{1}{\sqrt{\ell}}(Z_{1\star}, \frac{1}{2}(Z_{2\star} + Z_{2\star}^\top))$ is another optimal solution for (60). This proves the first part of the theorem.

For the second part, if there exists a symmetric optimal solution, then it completes the proof. Otherwise, from the first part, we know that there exists an optimal solution $(Z_{1\star}, Z_{2\star})$ with $Z_{1\star} = 0$, $Z_{2\star} = -Z_{2\star}^\top \in \mathbb{R}^{m\times m}$ and $\|(Z_{1\star}, Z_{2\star})\|_F = 1$ to (60). Let $i = \sqrt{-1}$. We have the following matrix decomposition:

$$Z_{2\star} = Q^\top \operatorname{diag}\left(\begin{bmatrix} 0 & -\omega_1 \\ \omega_1 & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & -\omega_k \\ \omega_k & 0 \end{bmatrix}, 0_r \right) Q,$$

where 0_r is a zero matrix with size $r \times r$, Q is an $m \times m$ orthogonal matrix, and $\omega_j > 0$ for $1 \le j \le k$. Let

$$\widetilde{Z}_{2\star} := Q^\top \operatorname{diag}\left(\begin{bmatrix} \omega_1 & 0 \\ 0 & \omega_1 \end{bmatrix}, \dots, \begin{bmatrix} \omega_k & 0 \\ 0 & \omega_k \end{bmatrix}, 0_r \right) Q$$

be a real symmetric matrix. Since $-\widetilde{Z}_{2\star} \preceq iZ_{2\star} \preceq \widetilde{Z}_{2\star}$, it is true that

$$\mathcal{L}(0,\widetilde{Z}_{2\star}) \succeq \mathcal{L}(0,iZ_{2\star}) \succeq \mathcal{L}(0,-\widetilde{Z}_{2\star}) = -\mathcal{L}(0,\widetilde{Z}_{2\star}).$$

Using the fact that $\|iZ_{2\star}\|_F = \|\widetilde{Z}_{2\star}\|_F$, we see that $\|(0,\widetilde{Z}_{2\star})\|_F = \|(0,iZ_{2\star})\|_F = \|(0,Z_{2\star})\|_F = 1$ and

$$\big\|\mathcal{L}(0,\widetilde{Z}_{2\star})\big\|_F \geq \big\|\mathcal{L}(0,iZ_{2\star})\big\|_F = \big\|\mathcal{L}(0,Z_{2\star})\big\|_F.$$

If $W_1 \succeq W_2 \succeq -W_1$, then $\|W_1\|_F \geq \|W_2\|_F$, which implies that $(0,\widetilde{Z}_{2\star})$ is a symmetric optimal solution to (60) (see [25, Lemma A.1]). This completes the proof. □

With the existence theory established above, it is interesting to note that the condition number $c(X)$ defined by (56) can be written as

$$\begin{aligned} c(X) &= \frac{1}{\kappa} \lim_{\delta\to 0^+} \sup_{\Omega_\delta} \frac{\|\mathcal{P}\Delta A + \mathcal{Q}\Delta B + \mathcal{M}\Delta C + \mathcal{N}\Delta D + \mathcal{H}\Delta S + \mathcal{T}\Delta R + \mathcal{L}_c^{-1}\Delta H\|_F}{\delta} \\ &= \frac{1}{\kappa} \max_{\delta_p>0} \frac{\|\mathcal{P}\Delta A + \mathcal{Q}\Delta B + \mathcal{M}\Delta C + \mathcal{N}\Delta D + \mathcal{H}\Delta S + \mathcal{T}\Delta R + \mathcal{L}_c^{-1}\Delta H\|_F}{\delta_p}. \end{aligned} \tag{63}$$

Note that the second equality in (63) is only an application of linearity of the norm. (For the proof, see Lemma A.1.) Observe further that the inverse operator $\mathcal{L}_c^{-1}$ of (4) satisfies

$$\big[\mathcal{L}_c^{-1}(W)\big]^\top = \mathcal{L}_c^{-1}\big(W^\top\big)$$

since $[\mathcal{L}_c(W)]^\top = \mathcal{L}_c(W^\top)$ for all $W \in \mathbb{C}^{n\times n}$. It follows that

$$\begin{aligned} &[\mathcal{T}\Delta R]^\top = \mathcal{T}\Delta R^\top, \qquad [\mathcal{P}\Delta A]^\top = \mathcal{P}\Delta A, \qquad [\mathcal{Q}\Delta B]^\top = \mathcal{Q}\Delta B, \\ &[\mathcal{M}\Delta C]^\top = \mathcal{M}\Delta C, \qquad [\mathcal{N}\Delta D]^\top = \mathcal{N}\Delta D, \qquad [\mathcal{H}\Delta S]^\top = \mathcal{H}\Delta S. \end{aligned}$$

Also, it is known that the inverse operator $\mathcal{L}_c^{-1}$ is positive [3, Corollary 3.8]. It follows that $\mathcal{T}$ is also a positive operator. Now, applying Theorem 5.1 to the operator $\mathcal{P}\Delta A + \mathcal{Q}\Delta B + \mathcal{M}\Delta C + \mathcal{N}\Delta D + \mathcal{H}\Delta S + \mathcal{T}\Delta R + \mathcal{L}_c^{-1}\Delta H$ in (63), we obtain the equality

$$\begin{aligned} &c(X) \\ &\quad = \frac{1}{\kappa} \max_{\widetilde{\Omega}} \frac{\|\kappa_A\mathcal{P}\Delta A + \kappa_B\mathcal{Q}\Delta B + \kappa_C\mathcal{M}\Delta C + \kappa_D\mathcal{N}\Delta D + \kappa_S\mathcal{H}\Delta S + \kappa_R\mathcal{T}\Delta R + \kappa_H\mathcal{L}_c^{-1}\Delta H\|_F}{\|(\Delta A,\Delta B,\Delta C,\Delta D,\Delta S,\Delta R,\Delta H)\|_F}, \end{aligned}$$

where the extended set $\widetilde{\Omega}$ is defined by

$$\begin{aligned} \widetilde{\Omega} = \big\{&(\Delta A,\Delta B,\Delta C,\Delta D,\Delta S,\Delta R,\Delta H) \in \mathbb{R}^{n\times n}\times\mathbb{R}^{n\times m}\times\mathbb{R}^{n\times n}\times\mathbb{R}^{n\times m} \\ &\times\mathbb{R}^{n\times n}\times\mathbb{R}^{n\times n}\times\mathbb{R}^{m\times m} \mid \big\|(\Delta A,\Delta B,\Delta C,\Delta D,\Delta S,\Delta R,\Delta H)\big\|_F > 0\big\}. \end{aligned}$$

On the other hand, observe that the matrix representation of the operation $\mathcal{L}_c$ in (4) can be written in terms of $L_c = I\otimes\Phi + \Phi^\top\otimes I + \Psi^\top\otimes\Psi$. Corresponding to (24a)-(24d) and

(58a)-(58b), we let

$$\begin{aligned}
M_c &= L_c^{-1}\big(I \otimes K^\top + \big(K^\top \otimes I\big)P_{n,n}\big),\\
N_c &= L_c^{-1}\big(F^\top \otimes K^\top + \big(K^\top \otimes F^\top\big)P_{m,n}\big),\\
T_c &= L_c^{-1}\big(F^\top \otimes F^\top\big),\\
H_c &= L_c^{-1}\big(F^\top \otimes I + \big(I \otimes F^\top\big)P_{m,n}\big),\\
P_c &= L_c^{-1}\big(I \otimes X + (X \otimes I)P_{n,n}\big),\\
Q_c &= L_c^{-1}\big(F^\top \otimes X + \big(X \otimes F^\top\big)P_{m,n}\big)
\end{aligned}$$

and

$$\mathcal{U} = \big(\kappa_A P_c, \kappa_B Q_c, \kappa_C M_c, \kappa_D N_c, \kappa_S H_c, \kappa_R T_c, \kappa_H L_c^{-1}\big).$$

It follows that

$$c(X) = \frac{1}{\kappa}\max_{V\in\widetilde{\Omega}} \frac{\|\mathcal{U}\operatorname{vec}(V)\|_2}{\|\operatorname{vec}(V)\|_2} = \frac{\|\mathcal{U}\|_2}{\kappa}.$$

Based on the above discussion, we have the following result.

Theorem 5.2 *The condition number $c(X)$ given by* (56) *has the explicit expression* $\frac{\|\mathcal{U}\|_2}{\kappa}$. *In particular, we have the relative condition number*

$$c_{\mathrm{rel}}(X) = \frac{\|(\|A\|_F P_c, \|B\|_F Q_c, \|C\|_F M_c, \|D\|_F N_c, \|S\|_F H_c, \|R\|_F T_c, \|H\|_F L_c^{-1})\|_2}{\|X\|_F}. \tag{64}$$

6 Numerical experiment

In this section we want to demonstrate the sharpness of perturbation bound (55) and its relationship with the relative condition number (64). Based on Newton's iteration [3], a numerical example, done with 2×2 coefficient matrices, is illustrated. The numerical algorithm is described in Algorithm 1. The corresponding stopping criterion is determined when the value of the *Normalized Residual* (NRes)

$$\mathrm{NRes} = \frac{\|\widetilde{\mathcal{P}}(\widetilde{X}) - \widetilde{\mathcal{S}}(\widetilde{X})\widetilde{\mathcal{Q}}(\widetilde{X})^{-1}\widetilde{\mathcal{S}}(\widetilde{X})^\top\|}{\|\widetilde{\mathcal{P}}(\widetilde{X})\| + \|\widetilde{\mathcal{S}}(\widetilde{X})\|\,\|\widetilde{\mathcal{Q}}(\widetilde{X})^{-1}\|\,\|\widetilde{\mathcal{S}}(\widetilde{X})^\top\|}$$

is less than or equal to a prescribed tolerance.

Example 1 Given a parameter $r = 10^{-m}$, for some $m > 0$, let the matrices A, B, C, D be defined by

$$A = \begin{bmatrix} -4 & 0 \\ 0 & -1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 0 & -\sqrt{r} \end{bmatrix}, \qquad C = I_2, \qquad D = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

and the matrices S, R, H be defined by

$$S = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad R = \begin{bmatrix} 3 & 0 \\ 0 & r \end{bmatrix}, \qquad H = \begin{bmatrix} -13/2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Algorithm 1: SARE $[\widetilde{X}] = \textsc{Sare}(\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}, \widetilde{R}, \widetilde{S}, \widetilde{H})$

Input: Matrices $\widetilde{A}, \widetilde{C} \in \mathbb{R}^{n\times n}$, $\widetilde{B}, \widetilde{D}, \widetilde{S} \in \mathbb{R}^{n\times m}$, $\widetilde{H} \in \mathcal{S}^n$, $\widetilde{R} \in \mathcal{S}^m$

Output: Matrix $\widetilde{X} \in \mathcal{S}^n$

begin

Choose $\widetilde{X}_0 \in \mathbb{R}^{n\times n}$;

for $i \leftarrow 1, \ldots$ **do**

$\widetilde{F}_k = -(\widetilde{R} + \widetilde{D}^\top \widetilde{X}_k \widetilde{D})^{-1}(\widetilde{B}^\top \widetilde{X}_k \widetilde{C} + \widetilde{S}^\top)$; $\widetilde{\Phi}_k = \widetilde{A} + \widetilde{B}\widetilde{F}_k$; $\widetilde{\Psi}_k = \widetilde{C} + \widetilde{D}\widetilde{F}_k$;

$\widetilde{\mathcal{R}}(\widetilde{X}_k)$ as defined by (7);

Updating $\widetilde{X}_{k+1}$ by solving

$\widetilde{\Phi}_k^\top \widetilde{X}_{k+1} + \widetilde{X}_{k+1}\widetilde{\Phi}_k + \widetilde{\Psi}_k^\top \widetilde{X}_{k+1}\widetilde{\Psi}_k = \widetilde{\Phi}_k^\top \widetilde{X}_k + \widetilde{X}_k \widetilde{\Phi}_k + \widetilde{\Psi}_k^\top \widetilde{X}_k \widetilde{\Psi}_k - \widetilde{\mathcal{R}}(\widetilde{X}_k)$;

end

end

Table 1 Relative errors and perturbation bounds

j	Relative error	$\frac{\xi_*}{\|X\|_F}$
5	6.28×10^{-5}	5.74×10^{-4}
6	6.34×10^{-6}	5.22×10^{-5}
7	1.14×10^{-6}	8.42×10^{-6}
8	1.61×10^{-7}	7.42×10^{-7}
9	1.05×10^{-8}	5.58×10^{-8}

It is easily seen that the unique stabilizing and maximal solution is

$$X = \begin{bmatrix} -1 & 0 \\ 0 & (\sqrt{5}-1)/2 \end{bmatrix}.$$

Let the perturbed coefficient matrices ΔA, ΔB, ΔC, ΔD, ΔS, ΔR and ΔH be generated using the MATLAB command randn with the weighted coefficient 10^{-j}. That is, the matrices ΔA, ΔB, ΔC, ΔD, ΔS, ΔR and ΔH are generated in forms of $\mathsf{randn}(2) \times 10^{-j}$, respectively. Since ΔR and ΔH are required to be symmetric, we need to fine-tune the perturbed matrices ΔR and ΔH by redefining ΔR and ΔH as $\Delta R + \Delta R^\top$ and $\Delta H + \Delta H^\top$, respectively. Now, let $(\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}, \widetilde{S}, \widetilde{R}, \widetilde{H}) = (A + \Delta A, B + \Delta B, C + \Delta C, D + \Delta D, S + \Delta S, R + \Delta R, H + \Delta H)$, which are coefficient matrices of SARE (5a)-(5b).

Firstly, we would like to evaluate the accuracy of the perturbation bound with the fixed parameter $r = 10^{-2}$, *i.e.*, $m = 2$, and different weighted coefficients, 10^{-j}, for $j = 5, \ldots, 9$. It can be seen from Table 1 that the values of the relative errors are closely bounded by our perturbation bounds of (55). In other words, (55) does provide a sharp upper bound of the relative errors of the stabilizing solution X.

Secondly, we want to investigate how ill-conditioned matrices affect the quantities of perturbation bounds. In this sense, the weighted coefficients are fixed to be 10^{-15}, *i.e.*, $j = 15$. The relationships among relative errors, perturbation bounds, and relative condition numbers are shown in Table 2. Due to the singularity of the matrix R caused by parameter r, the accuracy of the perturbation bounds is highly affected by the singularity. When the value of m increases, the perturbation bound is still tight to the relative error. Also, it can be seen that the number of accurate digits of the perturbation bounds is re-

Table 2 Relative errors, perturbation bounds and relative condition numbers

m	Relative error	$\frac{\xi_*}{\|X\|_F}$	$c_{\mathrm{rel}}(X)$
1	5.94×10^{-15}	1.64×10^{-14}	5.84×10^{1}
2	5.28×10^{-14}	9.47×10^{-14}	5.56×10^{2}
3	4.75×10^{-13}	7.85×10^{-13}	5.56×10^{3}
4	4.58×10^{-12}	7.40×10^{-12}	5.56×10^{4}
5	4.85×10^{-11}	7.26×10^{-11}	5.56×10^{5}
6	4.56×10^{-10}	7.22×10^{-10}	5.57×10^{6}
7	4.69×10^{-9}	8.61×10^{-9}	5.57×10^{7}

duced proportionally to the increase of the quantities of the relative condition numbers. In other words, if the accurate digits of the perturbation bound are added to the digits in the relative condition numbers, this number is almost equal to 16. (While using IEEE double-precision, the machine precision is around 2.2×10^{-16}.) This implies that the derived perturbation bound of (55) is fairly sharp.

7 Conclusion

While doing numerical computation, it is important in practice to have an accurate method for estimating the relative error and the condition number of the given problems. In this paper, we focus on providing a tight perturbation bound of the stabilizing solution to SARE (1a)-(1b) under small changes in the coefficient matrices. Also, some sufficient conditions are presented for the existence of the stabilizing solution to the perturbed SARE. The corresponding condition number of the stabilizing solution is provided in this work. We highlight and compare the practical performance of the derived perturbation bound and condition number through a numerical example. Numerical results show that our perturbation bound is very sensitive to the condition number of the stabilizing solution. As a consequence, they provide good measurement tools for the sensitivity analysis of SARE (1a)-(1b).

Appendix

We provide here a proof of the condition given by (63).

Lemma A.1 *Let* $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{M}$, $\mathcal{N}$, $\mathcal{H}$, $\mathcal{T}$, $\mathcal{L}_c^{-1}$ *be the operators defined by* (58a)-(58b), (24a)-(24d) *and* (4), *and let* Ω_δ, δ_p *be defined by* (57). *Then the following equality holds*:

$$\begin{aligned}&\lim_{\delta\to 0^+}\sup_{\Omega_\delta}\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta}\\&\quad=\max_{\delta_p>0}\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta_p}.\end{aligned}\tag{65}$$

Proof For any $\delta > 0$, $0 < \delta_p \le \delta$, we see that

$$\begin{aligned}&\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta}\\&\quad=\frac{\delta_p}{\delta}\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta_p}\\&\quad\le\max_{\delta_{\tilde p}>0}\frac{\|\mathcal{P}\Delta\widetilde A+\mathcal{Q}\Delta\widetilde B+\mathcal{M}\Delta\widetilde C+\mathcal{N}\Delta\widetilde D+\mathcal{H}\Delta\widetilde S+\mathcal{T}\Delta\widetilde R+\mathcal{L}_c^{-1}\Delta\widetilde H\|_F}{\delta_{\tilde p}},\end{aligned}$$

where $\delta_{\tilde{p}} = \|(\frac{\Delta\widetilde{A}}{\kappa_A}, \frac{\Delta\widetilde{B}}{\kappa_B}, \frac{\Delta\widetilde{C}}{\kappa_C}, \frac{\Delta\widetilde{D}}{\kappa_D}, \frac{\Delta\widetilde{S}}{\kappa_S}, \frac{\Delta\widetilde{R}}{\kappa_R}, \frac{\Delta\widetilde{H}}{\kappa_H})\|_F$. It follows that

$$\begin{aligned}&\lim_{\delta\to 0^+}\sup_{\Omega_\delta}\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta}\\&\quad\le\max_{\delta_p>0}\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta_p}.\end{aligned}\tag{66}$$

On the other hand, for any fixed $\delta > 0$, choose any perturbation matrices

$$(\Delta A_1, \Delta B_1, \Delta C_1, \Delta D_1, \Delta S_1, \Delta R_1, \Delta H_1) \in \Omega_\delta$$

and therefore

$$\left\|\left(\frac{\Delta A_1}{\kappa_A}, \frac{\Delta B_1}{\kappa_B}, \frac{\Delta C_1}{\kappa_C}, \frac{\Delta D_1}{\kappa_D}, \frac{\Delta S_1}{\kappa_S}, \frac{\Delta R_1}{\kappa_R}, \frac{\Delta H_1}{\kappa_H}\right)\right\|_F = \delta_{p_1} \le \delta.$$

It is true that $(\frac{\delta}{\delta_{p_1}}\Delta A_1, \frac{\delta}{\delta_{p_1}}\Delta B_1, \frac{\delta}{\delta_{p_1}}\Delta C_1, \frac{\delta}{\delta_{p_1}}\Delta D_1, \frac{\delta}{\delta_{p_1}}\Delta S_1, \frac{\delta}{\delta_{p_1}}\Delta R_1, \frac{\delta}{\delta_{p_1}}\Delta H_1) \in \Omega_\delta$ and this gives the fact that

$$\begin{aligned}&\lim_{\delta\to 0^+}\sup_{\Omega_\delta}\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta}\\&\quad\ge\frac{\|\mathcal{P}\frac{\delta}{\delta_{p_1}}\Delta A_1+\mathcal{Q}\frac{\delta}{\delta_{p_1}}\Delta B_1+\mathcal{M}\frac{\delta}{\delta_{p_1}}\Delta C_1+\mathcal{N}\frac{\delta}{\delta_{p_1}}\Delta D_1+\mathcal{H}\frac{\delta}{\delta_{p_1}}\Delta S_1+\mathcal{T}\frac{\delta}{\delta_{p_1}}\Delta R_1+\mathcal{L}_c^{-1}\frac{\delta}{\delta_{p_1}}\Delta H_1\|_F}{\delta}\\&\quad=\frac{\|\mathcal{P}\Delta A_1+\mathcal{Q}\Delta B_1+\mathcal{M}\Delta C_1+\mathcal{N}\Delta D_1+\mathcal{H}\Delta S_1+\mathcal{T}\Delta R_1+\mathcal{L}_c^{-1}\Delta H_1)\|_F}{\delta_{p_1}}.\end{aligned}$$

Hence

$$\begin{aligned}&\lim_{\delta\to 0^+}\sup_{\Omega_\delta}\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta}\\&\quad\ge\max_{\delta_p>0}\frac{\|\mathcal{P}\Delta A+\mathcal{Q}\Delta B+\mathcal{M}\Delta C+\mathcal{N}\Delta D+\mathcal{H}\Delta S+\mathcal{T}\Delta R+\mathcal{L}_c^{-1}\Delta H\|_F}{\delta_p}.\end{aligned}\tag{67}$$

Comparison of (66) and (67) gives (65). □

Competing interests
The authors declare that there is no conflict of interests regarding the publication of this article.

Authors' contributions
All authors contributed equally and significantly in writing this paper. All authors read and approved the final manuscript.

Author details
[1]Center for General Education, National Formosa University, Huwei 632, Taiwan. [2]Department of Mathematics, National Taiwan Normal University, Taipei 116, Taiwan. [3]Department of Mathematics, National Chung Cheng University, Chia-Yi 621, Taiwan.

Acknowledgements
The authors wish to thank the editor and two anonymous referees for many interesting and valuable suggestions on the manuscript. This research work is partially supported by the National Science Council and the National Center for Theoretical Sciences in Taiwan. The first author was supported by the National Science Council of Taiwan under Grant NSC 102-2115-M-150-002. The second author was supported by the National Science Council of Taiwan under Grant NSC 102-2115-M-003-009. The third author was supported by the National Science Council of Taiwan under Grant NSC 101-2115-M-194-007-MY3.

Received: 16 September 2013 Accepted: 14 November 2013 Published: 11 Dec 2013

References

1. Rami, MA, Zhou, XY: Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls. IEEE Trans. Autom. Control **45**(6), 1131-1143 (2000). doi:10.1109/9.863597
2. El Bouhtouri, A, Hinrichsen, D, Pritchard, AJ: On the disturbance attenuation problem for a wide class of time invariant linear stochastic systems. Stoch. Stoch. Rep. **65**(3-4), 255-297 (1999)
3. Damm, T, Hinrichsen, D: Newton's method for a rational matrix equation occurring in stochastic control. In: Proceedings of the Eighth Conference of the International Linear Algebra Society (Barcelona, 1999), vol. 332/334, pp. 81-109 (2001)
4. Hinrichsen, D, Pritchard, AJ: Stochastic H_∞. SIAM J. Control Optim. **36**, 1504-1538 (1998)
5. Lancaster, P, Rodman, L: Algebraic Riccati Equations. Oxford Science Publications. Clarendon, New York (1995)
6. Mehrmann, VL: The Autonomous Linear Quadratic Control Problem: Theory and Numerical Solution. Lecture Notes in Control and Information Sciences, vol. 163. Springer, Berlin (1991). doi:10.1007/BFb0039443
7. Zhou, K, Doyle, JC, Glover, K: Robust and Optimal Control. Prentice Hall, Upper Saddle River (1996)
8. Benner, P, Laub, AJ, Mehrmann, V: A collection of benchmark examples for the numerical solution of algebraic Riccati equations I: continuous-time case. Technical Report SPC 95_22, Fakultät für Mathematik, TU Chemnitz-Zwickau, 09107 Chemnitz, FRG. http://www.tu-chemnitz.de/sfb393/spc95pr.html (1995)
9. Benner, P, Laub, AJ, Mehrmann, V: A collection of benchmark examples for the numerical solution of algebraic Riccati equations II: discrete-time case. Technical Report SPC 95_23, Fakultät für Mathematik, TU Chemnitz-Zwickau, 09107 Chemnitz, FRG. http://www.tu-chemnitz.de/sfb393/spc95pr.html (1995)
10. Sima, V: Algorithms for Linear-Quadratic Optimization. Monographs and Textbooks in Pure and Applied Mathematics, vol. 200. Dekker, New York (1996)
11. Laub, AJ: A Schur method for solving algebraic Riccati equations. IEEE Trans. Autom. Control **24**(6), 913-921 (1979)
12. Ammar, G, Benner, P, Mehrmann, V: A multishift algorithm for the numerical solution of algebraic Riccati equations. Electron. Trans. Numer. Anal. **1**, 33-48 (1993)
13. Ammar, G, Mehrmann, V: On Hamiltonian and symplectic Hessenberg forms. Linear Algebra Appl. **149**, 55-72 (1991)
14. Benner, P, Mehrmann, V, Xu, H: A new method for computing the stable invariant subspace of a real Hamiltonian matrix. J. Comput. Appl. Math. **86**, 17-43 (1997)
15. Benner, P, Mehrmann, V, Xu, H: A numerically stable, structure preserving method for computing the eigenvalues of real Hamiltonian or symplectic pencils. Numer. Math. **78**(3), 329-358 (1998)
16. Bunse-Gerstner, A, Byers, R, Mehrmann, V: A chart of numerical methods for structured eigenvalue problems. SIAM J. Matrix Anal. Appl. **13**, 419-453 (1992)
17. Bunse-Gerstner, A, Mehrmann, V: A symplectic QR like algorithm for the solution of the real algebraic Riccati equation. IEEE Trans. Autom. Control **31**(12), 1104-1113 (1986)
18. Mehrmann, V: A step toward a unified treatment of continuous and discrete time control problems. Linear Algebra Appl. **241-243**, 749-779 (1996)
19. Byers, R: Solving the algebraic Riccati equation with the matrix sign function. Linear Algebra Appl. **85**, 267-279 (1987)
20. Benner, P: Contributions to the numerical solutions of algebraic Riccati equations and related eigenvalue problems. PhD thesis, Fakultät für Mathematik, TU Chemnitz-Zwickau, Chemnitz, Germany (1997)
21. Chu, EK-W, Fan, H-Y, Lin, W-W: A structure-preserving doubling algorithm for continuous-time algebraic Riccati equations. Linear Algebra Appl. **396**, 55-80 (2005)
22. Chu, EK-W, Fan, H-Y, Lin, W-W, Wang, C-S: Structure-preserving algorithms for periodic discrete-time algebraic Riccati equations. Int. J. Control **77**(8), 767-788 (2004)
23. Chiang, C-Y, Fan, H-Y: Residual bounds of the stochastic algebraic Riccati equation. Appl. Numer. Math. **63**, 78-87 (2013)
24. Konstantinov, M, Gu, D-W, Mehrmann, V, Petkov, P: Perturbation Theory for Matrix Equations. Studies in Computational Mathematics, vol. 9. North-Holland, Amsterdam (2003)
25. Sun, J-G: Perturbation theory for algebraic Riccati equations. SIAM J. Matrix Anal. Appl. **19**(1), 39-65 (1998). doi:10.1137/S0895479895291303
26. Sun, J-g: Sensitivity analysis of the discrete-time algebraic Riccati equation. In: Proceedings of the Sixth Conference of the International Linear Algebra Society (Chemnitz, 1996), vol. 275/276, pp. 595-615 (1998). doi:10.1016/S0024-3795(97)10017-9
27. Sun, J-g: Residual bounds of approximate solutions of the algebraic Riccati equation. Numer. Math. **76**(2), 249-263 (1997). doi:10.1007/s002110050262
28. Sun, J-g: Residual bounds of approximate solutions of the discrete-time algebraic Riccati equation. Numer. Math. **78**(3), 463-478 (1998). doi:10.1007/s002110050321
29. Riedel, KS: A Sherman-Morrison-Woodbury identity for rank augmenting matrices with application to centering. SIAM J. Matrix Anal. Appl. **13**(2), 659-662 (1992). doi:10.1137/0613040
30. Stewart, GW, Sun, JG: Matrix Perturbation Theory. Computer Science and Scientific Computing. Academic Press, Boston (1990)
31. Ortega, JM, Rheinboldt, WC: Iterative Solution of Nonlinear Equations in Several Variables. Classics in Applied Mathematics, vol. 30. SIAM, Philadelphia (2000). Reprint of the 1970 original
32. Rice, JR: A theory of condition. SIAM J. Numer. Anal. **3**, 287-310 (1966)
33. Sun, J-g: Condition numbers of algebraic Riccati equations in the Frobenius norm. Linear Algebra Appl. **350**, 237-261 (2002). doi:10.1016/S0024-3795(02)00294-X
34. Xu, S: Matrix Computation in Control Theory. Higher Education Press, Beijing (2010) (In Chinese)

10.1186/1029-242X-2013-580
Cite this article as: Chiang et al.: **Perturbation analysis of the stochastic algebraic Riccati equation.** *Journal of Inequalities and Applications* 2013, 2013:580